Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A threshold fuzzy entropy based feature selection for medical database classification

P. Jaganathan, R. Kuppuchamy*

Department of Computer Applications, PSNA College of Engineering and Technology, Dindigul 624622, Tamilnadu, India

ARTICLE INFO

Article history:
Received 14 February 2012
Accepted 19 October 2013

Keywords:
Fuzzy entropy
Feature selection
Medical database
Classification

ABSTRACT

Feature selection is one of the most common and critical tasks in database classification. It reduces the computational cost by removing insignificant features. Consequently, this makes the diagnosis process accurate and comprehensible. This paper presents the measurement of feature relevance based on fuzzy entropy, tested with a Radial Basis Function Network classifier for a medical database classification. Three feature selection strategies are devised to obtain the valuable subset of relevant features. Five benchmarked datasets, which are available in the UCI Machine Learning Repository, have been used in this work. The classification accuracy shows that the proposed method is capable of producing good results with fewer features than the original datasets.

1. Introduction

The revolution in database technologies has resulted in an increase of data accumulation in many areas, such as financial, marketing and the biological and medical sciences. It has become crucial to locate hidden information by scrutinizing these data effectively. Data mining techniques have been discussed widely and applied successfully in the areas of medical research, scientific analysis and business applications. Recently, the absorption of data mining techniques in medical diagnosis has provided new insights in a large number of medical applications [1–3]. Feature selection has many advantages such as shortening the number of measurements, reducing the execution time and improving transparency and compactness of the suggested diagnosis.

Feature selection is the process of selecting a subset of ‘d’ features of the set D, such that $d \leq D$. The primary purpose of feature selection is to reduce the computational cost and to improve the performance of the learning algorithm. Feature selection algorithms deal with different evaluation criteria and generally, are classified into filter and wrapper models [4–6]. The filter model evaluates the general characteristics of the training data to select a feature subset without relation to any learning algorithms; thus, it is computationally economical. Nevertheless, it carries the risk of selecting subsets of features that may not be relevant. The wrapper model requires a pre-determined induction algorithm, which assesses the performance of the features that are chosen. The selected features are related significantly to the choice of the classifier and do not generalize to other classifiers. However, this tends to be computationally expensive. Therefore, the filter and wrapper models could complement each other; wrapper models provide better accuracy, whereas filter models search the feature space efficiently.

This paper proposes a filter-based feature subset selection based on fuzzy entropy measures and presents the different selection strategies for handling medical database classification. The proposed method is evaluated using a Radial Basis Function (RBF) cilassification algorithm for the given benchmark datasets.

The remainder of the paper is organized as follows. A brief literature review is presented in Section 2. The methods and materials employed in this study are introduced in Section 3. The observational results and discussion are reported in Sections 4 and 5. Section 6 provides the conclusions and directions for future research.

2. Literature review

Recently, a number of researchers have focused on several feature selection methods and most of them have reported their good performance in database classification.

Battiti [7] proposes a method called Mutual-Information-based Feature Selection (MIFS), in which the selection criterion is based on maximizing the mutual information between candidate features and the class variables, and minimizing the redundancy between candidate features and the selected features. Hanchuan et al. [8] follow a similar technique to MIFS, which has been called the minimal-redundancy-maximal-relevance (mRMR) criterion.

* Corresponding author. Tel.: +91 9842033475.
E-mail addresses: jaganathodc@gmail.com (P. Jaganathan), rkuppuchamy@yahoo.com (R. Kuppuchamy).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.10.016

It eliminates the manually tuned parameter with cardinality of the features already selected. Pablo et al. [9] present a Normalized Mutual Information Feature Selection algorithm. The mutual information among features should be divided by the minimum value of their entropies in order to produce a normalized value, which is to be measured by the redundant term. Yu and Liu [10] developed a correction-based method for relevance and redundancy analysis and then removed redundant features using the Markov Blanket method.

In addition, feature selection methods are analyzed by a number of techniques. Abdel-Aal [1] developed a novel technique for feature ranking and selection with the group method of data handling. Feature reduction of more than 50% could be achieved and improved in the classification performance. Sahan et al. [11] built a new hybrid machine learning method for a fuzzy-artificial immune system with a k-nearest neighbor algorithm to solve medical diagnosis problems, which demonstrated good results. Jaganathan et al. [12] applied a new improved quickreduct algorithm, which is a variant of quickreduct for feature selection and tested it on a classification algorithm called Ant Miner. Sivagaminathan et al. [13] proposed a hybrid method combining Ant Colony Optimization and Artificial Neural Networks (ANNs) to deal with feature selection, which produced promising results. Lin et al. [14] proposed a Simulated Annealing approach for parameter setting in Support Vector Machines, which is compared with a grid search parameter setting and was found to produce higher classification accuracy.

Lin et al. [15] applied a Particle-Swarm-Optimization-based approach to search for appropriate parameter values for a back-propagation network to select the most valuable subset of features to improve classification accuracy. Unler et al. [16] developed a modified discrete particle swarm optimization algorithm for the feature selection problem and compared it with tabu and scatter search algorithms to demonstrate its effectiveness.

Chang et al. [17] introduced a hybrid model for integrating a case-based reasoning approach with a particle swarm optimization model for feature subset selection in medical database classification. Salamo et al. [18] evaluated a number of measures for estimating feature relevance based on rough set theory and also proposed three strategies for feature selection in a Case Based Reasoning classifier. Qasem et al. [19] applied a time variant multi-objective particle swarm optimization to an RBF Network for diagnosing medical diseases.

This paper describes in detail how to combine the relevance measures and feature subset selection strategies.

3. Methods and materials

3.1. Fuzzy entropy-based relevance measure

In information theory, the Shannon entropy measure is generally used to characterize the impurity of a collection of samples. Assuming X as a discrete random variable with a finite set of n elements, where $X=\{x_1, x_2, x_3,\ldots,x_n\}$, then if an element x_i occurs with probability $p(x_i)$ [20], the entropy $H(X)$ of X is defined as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{1}$$

where n denotes the number of elements.

An extension of Shannon entropy with fuzzy sets, which is used to support the evaluation of entropies, is called fuzzy entropy. It was introduced in 1972 [21], after which a number of modifications were introduced to the original fuzzy entropy method [22,23].

The proposed fuzzy entropy method is based on the utilization of the Fuzzy C-Means Clustering algorithm (FCM), which is used to construct the membership function of all features. The data may belong to two or more clusters simultaneously and the belonging of a data point to the clusters is governed by the membership values [24]. Similar data points are placed in the same cluster and dissimilar data points normally belong to different clusters. The membership values of the data points are reorganized iteratively to reduce the dissimilarity. The Euclidean distance is used to measure the dissimilarity of two data points.

The FCM algorithm is explained as follows [25].

Step 1: assume the number of clusters (C), where $2 \le C \le N$, C – number of clusters and N – number of data points
Step 2: calculate the jth cluster center C_j using the following expression

$$C_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{g} x_{ij}}{\sum_{i=1}^{N} \mu_{ij}^{g}} \tag{2}$$

where $g \ge 1$ is the fuzziness coefficient and μ_{ij} is the degree of membership for the ith data point x_i in cluster j.
Step 3: calculate the Euclidean distance between the ith data point and the jth cluster center as follows:

$$d_{ij} = |C_j - x_i| \tag{3}$$

Step 4: update the fuzzy membership values according to d_{ij}. If $d_{ij} \ge 0$, then

$$\mu = \frac{1}{\sum_{m=1}^{C} \left(\frac{d_{ij}}{d_{im}}\right)^{\frac{2}{g-1}}} \tag{4}$$

If $d=0$, then the data point coincides with the jth cluster center (C) and it will have the full membership value, i.e., $\mu_{ij}=1.0$
Step 5: repeat Steps 2–4 until the changes in $[\mu]$ are less than some pre-specified values.

The FCM algorithm computes the membership of each sample in all clusters and then normalizes it. This procedure is applied for each feature. The summation of membership of feature 'x' in class 'c', divided by the membership of feature 'x' in all 'C' classes, is termed the class degree $CD_c(\tilde{A})$ [26], which is given as

$$CD_c(\tilde{A}) = \frac{\sum_{x \in c} \mu_{\tilde{A}}(x)}{\sum_{x \in C} \mu_{\tilde{A}}(x)} \tag{5}$$

where $\mu_{\tilde{A}}$ denotes the membership function of the fuzzy set and $\mu_{\tilde{A}}(x_i)$ denotes the membership grade of x belonging to the fuzzy set $\tilde{A}$.

The fuzzy entropy $FE_c(\tilde{A})$ of class 'c' is defined as

$$FE_c(\tilde{A}) = -CD_c(\tilde{A})\log_2 CD_c(\tilde{A}) \tag{6}$$

The fuzzy entropy FE $(\tilde{A})$ of a fuzzy set X is defined as follows:

$$FE(\tilde{A}) = \sum_{c \in C} FE_c(\tilde{A}) \tag{7}$$

The probability $p(x_i)$ of Shannon's entropy is measured by the number of occurring elements. In contrast, the class degree $CD_c(\tilde{A})$ in fuzzy entropy is measured by the membership values of the occurring elements. Suppose the elements are divided into a number of intervals: I_1, I_2, then the Shannon entropy of interval I_1 is equal to the interval I_2. However, fuzzy entropy can indicate that interval I_1 is distinguishable from interval I_2. The difference between Shannon's entropy and the proposed fuzzy entropy is illustrated in Fig. 1, where the symbols 'M' and 'F' denote male and female samples, respectively. The computation of Shannon's

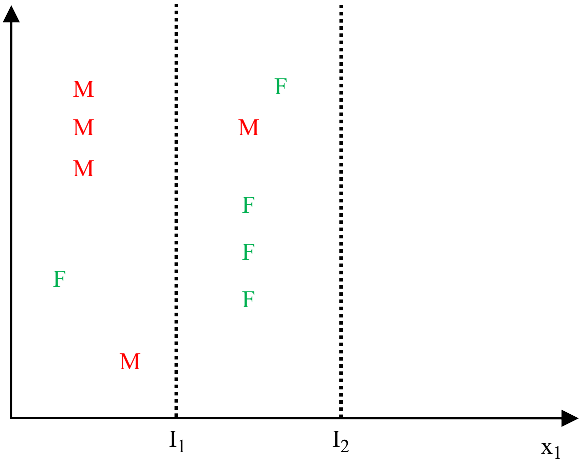

Fig. 1. Distribution of samples.

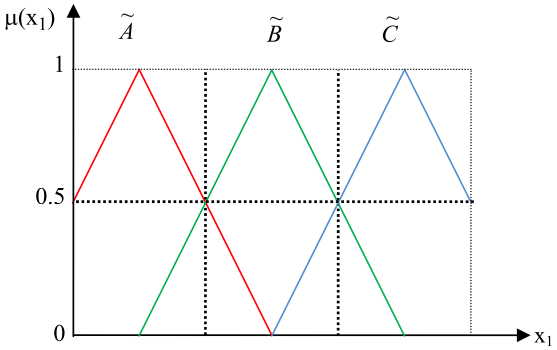

Fig. 2. A feature x_1 with fuzzy sets $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$.

entropy in the intervals I_1 and I_2 is as follows:

$$H(I_1) = -(p(M)\log_2 p(M) + p(F)\log_2 p(F))$$

$$H(I_1) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) \cong 0.72$$

$$H(I_2) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) \cong 0.72$$

The fuzzy entropy of intervals I_1 and I_2 for the corresponding fuzzy sets $\tilde{A}$ is shown in Fig. 2.

$$\sum_{x \in M} \mu_{\tilde{A}}(x) = 0.75 + 1 + 1 + 1 = 3.75$$

$$\sum_{x \in F} \mu_{\tilde{A}}(x) = 0.75$$

Based on Eq. (5), the class degrees of 'M' and 'F' are

$$CD_M(\tilde{A}) = \frac{3.75}{3.75 + 0.75} = 0.833$$

$$CD_F(\tilde{A}) = \frac{0.75}{3.75 + 0.75} = 0.167$$

The fuzzy entropy FE $(\tilde{A})$ of a fuzzy set $\tilde{A}$ is calculated as follows:

$$FE(\tilde{A}) = FE_M(\tilde{A}) + FE_F(\tilde{A})$$

$$FE(\tilde{A}) = -(CD_M(\tilde{A})\log_2 CD_M(\tilde{A}) + CD_F(\tilde{A})\log_2 CD_F(\tilde{A}))$$

$$FE(\tilde{A}) = -(0.833\log_2 0.833 + 0.167\log_2 0.167) \cong 0.651$$

Similarly, the fuzzy entropy of the fuzzy set $\tilde{B}$ is as follows:

$$\sum_{x \in M} \mu_{\tilde{B}}(x) = 1 + 0.25 = 1.25$$

$$\sum_{x \in F} \mu_{\tilde{B}}(x) = 1 + 1 + 1 + 0.75 = 3.75$$

The class degree of 'M' is 1.25/(1.25+3.75)=0.25 and 3.75/(1.25+3.75)=0.75.

The fuzzy entropy $FE(\tilde{B})$ is

$$FE(\tilde{B}) = -(0.25\log_2 0.25 + 0.75\log_2 0.75) \cong 0.811$$

From the above results, Shannon's entropy of interval I_1 is equal to interval I_2. Nonetheless, the fuzzy entropy indicates that interval I_1 is distinct from interval I_2. The highest fuzzy entropy value of the feature is regarded as the most informative one.

3.2. Feature selection strategies

This subsection explains three different criteria for the feature selection process. The features are regulated with respect to decreasing values of the fuzzy entropy. A feature in the first position is the most relevant and the one in the last position is the least relevant in the resulting rank vector. The framework of feature selection is depicted in Fig. 3.

Mean Selection (MS) Strategy: a feature $f \in F$ is selected if it satisfies the following condition:

$$\sigma(f) \geq \sum_{f \in F} \frac{\sigma(f)}{|F|} \tag{8}$$

where $\sigma(f)$ is the relevance value of the features, which is selected if it is greater than or equivalent to the mean of the relevant values. This strategy will be useful in examining the suitability of the fuzzy entropy relevance measure.

Half Selection (HS) Strategy: the half selection strategy aims to reduce feature dimensionality to select approximately 50% of the features in the domain. The feature $f \in F$ is selected if it satisfies the following condition:

$$P_a \geq \frac{|F|}{2} \tag{9}$$

where P_a is the position of the feature in the rank vector. It represents the selected features having a relevance value higher than a given threshold, which is calculated as $|F|/2$. This strategy does produce great reductions, close to 50%. At the same time, some of the selected features are irrelevant despite them passing the threshold. Similarly, some of the omitted features may also be relevant despite them not being selected. This suggests that a new feature selection strategy must be based on the relevance value of each feature instead of a predefined number of features that are to be reduced. The last feature selection strategy described below has a relatively smaller number of features but at the same time, it retains the most relevant.

Neural Network for Threshold Selection (NNTS): an ANN [27] is one of the well-known machine learning techniques and it can be used in a variety of applications in data mining [28]. The ANN provides a variety of feedforward networks that are generally called backprogagation networks. It possesses a number of interconnected layers that consist of an input layer, a hidden layer and an output layer. The fuzzy entropy value of each feature is an initial value for each node in the input layer. The value from the input layer to the output layer is achieved by hidden layers using weights and activation functions. A sigmoid function is used as an activation function and a learning rate coefficient determines the size of weight adjustments made at the each iteration. An output layer is used to represent an output value. The output value can be considered as a threshold value of the given fuzzy entropy.

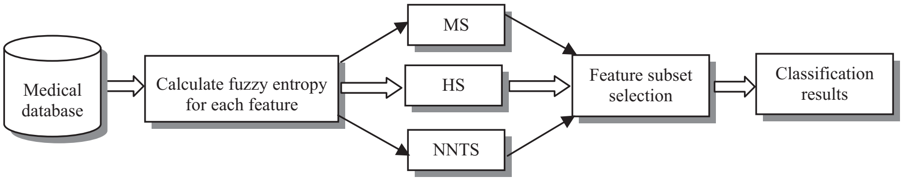

Fig. 3. Framework for feature selection.

Table 1
Details of datasets used.

Dataset	No. of instances	No. of features	No. of classes
Wisconsin Breast Cancer	699	9	2
Pima Indians Diabetes	768	8	2
Heart-Statlog	270	13	2
Hepatitis	155	19	2
Cleveland Heart Disease	296	13	5

The features to be selected have relevance values higher than or equal to a threshold value.

$$\sigma(f) \geq \text{threshold value} \tag{10}$$

3.3. Dataset used

The performance of the proposed method is evaluated using five benchmark datasets: Wisconsin Breast Cancer, Pima Indians Diabetes, Heart-Statlog, Hepatitis and Cleveland Heart Disease, which are available from the UCI Machine Learning Repository [29]. Table 1 summarizes the number of features, instances and classes for each dataset used in this study.

3.3.1. Wisconsin breast cancer

The dataset was collected by Dr. William H. Wolberg (1989–1991) at the University of Wisconsin–Madison Hospitals. It contains 699 instances characterized by nine features: (1) Clump Thickness, (2) Uniformity of Cell Size, (3) Uniformity of Cell Shape, (4) Marginal Adhesion, (5) Single Epithelial Cell Size, (6) Bare Nuclei, (7) Bland Chromatin, (8) Normal Nucleoli and (9) Mitoses, which are used to predict benign or malignant growths. In this dataset, 241 (34.5%) instances are malignant and 458 (65.5%) instances are benign.

3.3.2. Pima Indians diabetes

The dataset is available at the National Institute of Diabetes and Digestive and Kidney Diseases. It contains 768 instances described by eight features used to predict the presence or absence of diabetes. The features are as follows: (1) number of pregnancies, (2) plasma glucose concentration, (3) diastolic blood pressure, (4) tricep skin fold thickness, (5) serum insulin, (6) body mass index, (7) diabetes pedigree function and (8) age in years.

3.3.3. Heart-Statlog

The dataset is based on data from the Cleveland Clinic Foundation and it contains 270 instances belonging to two classes: the presence or absence of heart disease. It is described by 13 features (age, sex, chest, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic, maximum heart rate, exercise induced angina, oldpeak, slope, number of major vessels and thal).

3.3.4. Hepatitis

The dataset is obtained from the Carnegie–Mellon University and it contains 155 instances belonging to two classes: live or die. There are 19 features (age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver film, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, SGOT, albumin, protime and histology).

3.3.5. Cleveland heart disease

The dataset was collected from the Cleveland Clinic Foundation and contains about 296 instances, each having 13 features (originally 76 raw features), which are used to infer the presence (values 1, 2, 3, 4) or absence (value 0) of heart disease. The features are (1) age, (2) sex, (3) chest pain type, (4) resting blood pressure, (5) cholesterol, (6) fasting blood sugar, (7) resting electrocardiographic results, (8) maximum heart rate, (9) exercise induced angina, (10) depression induced by exercise relative to segment, (11) slope of peak exercise, (12) number of major vessels and (13) thal.

4. Results

The classification performance of the proposed feature selection method is measured using an RBF Network classifier. An RBF network is a type of ANN, which is a simpler network structure with better approximation capabilities. It is implemented with the WEKA (*Waikato Environment for Knowledge Analysis*) workbench [30]. The performances of the proposed methods are evaluated using 10-fold cross validation [31]. All datasets are split into 10 subsets of approximately equal size. Randomly, one dataset is used for testing and the remainder are used for training. The same procedure is repeated 10 times and the mean classification accuracy is computed. Tables 2–6 present the results reported for each dataset. The rows represent the results for the original features and the results for the reduced features with MS, HS and NNTS strategies. The columns represent the names of the feature selection method, classification accuracy, sensitivity, specificity, number of features selected, percentage of features selected, Area Under Receiver Operator Characteristic (AUROC) and the list of selected features. The proposed method was implemented with Matlab2007b and all experiments were performed on 3.2 GHz Pentium IV machines with 512 MB memory, running Windows XP.

4.1. Breast cancer dataset

In Table 2, the results are reported for different feature selection methods for the breast cancer dataset. When the dataset is classified using the original features, classification accuracy of 95.85%, sensitivity of 0.92 and a specificity of 0.98 are obtained. When the MS and HS strategies are applied, accuracies of 95.99% and 96.71% are obtained, respectively. When the NNTS strategy is applied, the accuracy is enhanced significantly to 97.28%. The best sensitivity of 0.94 is achieved with both the HS and NNTS strategies and the specificity of 0.99 is achieved with the NNTS strategy. The highest accuracy is reported for this dataset when the NNTS strategy is employed.

Table 2
Classification results with Breast Cancer dataset.

Method	Accuracy	Sensitivity	Specificity	Dimension	Features selected (%)	AUROC	Selected features
None	95.85	0.92	0.98	9	100.0	0.98	All
MS	95.99	0.93	0.97	4	44.4	0.98	2,3,6,7
HS	96.71	0.94	0.98	5	55.6	0.98	2,3,6,7,8
NNTS	97.28	0.94	0.99	7	77.8	0.98	1,2,3,5,6,7,8

Table 3
Classification results with Pima Indians Diabetes dataset.

Method	Accuracy	Sensitivity	Specificity	Dimension	Features selected (%)	AUROC	Selected features
None	73.83	0.65	0.78	8	100	0.79	All
MS	76.04	0.71	0.78	3	37.5	0.81	2,6,8
HS	75.91	0.69	0.79	4	50.0	0.80	1,2,6,8
NNTS	76.04	0.71	0.78	3	37.5	0.81	2,6,8

Table 4
Classification results with Heart-Statlog dataset.

Method	Accuracy	Sensitivity	Specificity	Dimension	Features selected (%)	AUROC	Selected features
None	84.44	0.82	0.86	13	100	0.89	All
MS	84.44	0.85	0.84	6	46.2	0.89	3,8,9,11,12,13
HS	84.81	0.85	0.84	7	53.8	0.89	3,8,9,10,11,12,13
NNTS	85.19	0.85	0.86	4	30.8	0.89	3,11,12,13

Table 5
Classification results with Hepatitis dataset.

Method	Accuracy	Sensitivity	Specificity	Dimension	Features selected (%)	AUROC	Selected features
None	84.52	0.90	0.63	19	100	0.81	All
MS	82.58	0.87	0.60	8	42.1	0.85	5,6,1,12,13,14,17,19
HS	85.16	0.90	0.66	10	52.6	0.86	2,5,6,10,11,12,13,14,17,19
NNTS	85.16	0.90	0.66	10	52.6	0.86	2,5,6,10,11,12,13,14,17,19

Table 6
Classification results with Cleveland Heart Disease dataset.

Method	Accuracy	Sensitivity	Specificity	Dimension	Features selected (%)	Selected features
None	83.82	0.83	0.85	13	100	All
MS	81.75	0.82	0.82	6	46.2	3,8,9,11,12,13
HS	83.44	0.84	0.83	7	53.8	3,8,9,10,11,12,13
NNTS	84.46	0.82	0.82	3	23.1	3,12,13

4.2. *Pima Indians diabetes dataset*

The performance of the different feature selection methods for the Pima Indians Diabetes dataset is shown in Table 3. As can be seen, both the MS and the NNTS strategies achieved best accuracy of 76.04% and a sensitivity of 0.71 with three features. Similarly, the specificity of 0.79 was obtained with the HS strategy. Therefore, there is a significant improvement in the performance of the feature selection algorithms with reduced features than with original features. The accuracy is enhanced from 73.83% to 76.04% with only three features, against the eight features of the original dataset.

4.3. *Heart-Statlog dataset*

Table 4 depicts the results obtained by the different feature selection methods with the Heart-Statlog dataset. A classification accuracy of 84.44% is achieved with all the 13 features of the dataset. The MS, HS and NNTS strategies reduced the number of features to 6, 7 and 4, respectively with classification accuracies of 84.44%, 84.81% and 85.19% obtained with the reduced features. It is clear from the above that a reduction in the number of features not only maintains the same classification accuracy as in the case of the MS strategy but also increases it, as in the case of the HS and NNTS strategies. Sensitivity of 0.85 is obtained for all of the three feature selection strategies compared with the sensitivity of 0.82 obtained for all the features. The specificity of 0.84 is obtained for the MS and HS strategies; however, the NNTS strategy yields a specificity of 0.86, as in the case of the specificity obtained from all the features.

4.4. *Hepatitis dataset*

The results for the hepatitis dataset are shown in Table 5. It can be seen that the best result is achieved with the HS and NNTS strategies with an accuracy of 85.16%, sensitivity of 0.90 and specificity of 0.66. The results with the original features provided an accuracy of 84.52%, sensitivity of 0.90 and specificity of 0.63, as tabulated. It is observed that only 10 features are needed to produce the best accuracy, as in the case of the HS and NNTS strategies, compared with those obtained with all the features.

4.5. Cleveland heart disease dataset

The results for the Cleveland Heart Disease dataset can be seen in Table 6. As we are now dealing with a five-class classification problem instead of a binary-class classification problem, the AUROC is omitted. A classification accuracy of 83.82%, sensitivity of 0.83 and a specificity of 0.85 are achieved with the original features of the dataset. The best accuracy of 84.46% is reached with the NNTS strategy. The MS strategy and HS strategy do not work well with this dataset and produce accuracies of only 81.75%, 83.44%, respectively. The best accuracy is seen with only three features, compared with the 13 features of the original dataset.

5. Discussion

The proposed feature selection methods have reduced the number of features instead of using all the features to perform the classification. The percentage of features selected (X-axis) in each feature selection strategy is plotted against the number of features (Y-axis) in the dataset in Figs. 4–6. Fig. 4 shows the results for the MS strategy. The lowest percentage of features selected is 37.5% for the Pima Indians Diabetes dataset and the highest is 46.2% for the Heart-Statlog and Cleveland Heart datasets. It is observed that this strategy has selected less than 50% of the features in each dataset. In Fig. 5, the HS strategy has selected 50% of the features in all datasets. The results obtained for the NNTS strategy (Fig. 6) show that 77.8% of the features of the Breast Cancer dataset have produced the best classification accuracy of 97.28%. For the Pima Indians Diabetes and Heart-Statlog datasets, 37.5% and 30.8% of features are selected. Similarly, for the Cleveland Heart Disease and hepatitis datasets, 23.1% and 52.6% of the original features are obtained.

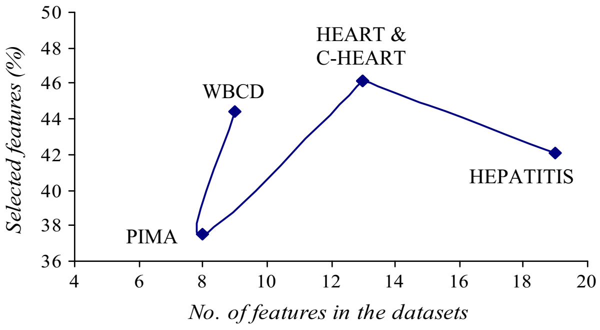

Fig. 4. Feature selection in terms of MS strategy.

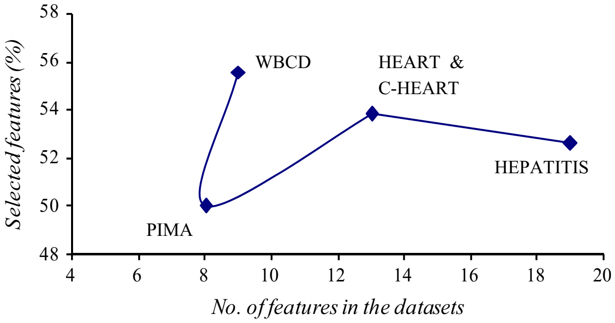

Fig. 5. Feature selection in terms of HS strategy.

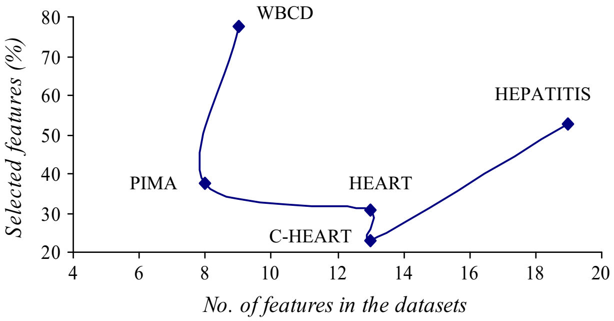

Fig. 6. Feature selection in terms of NNTS strategy.

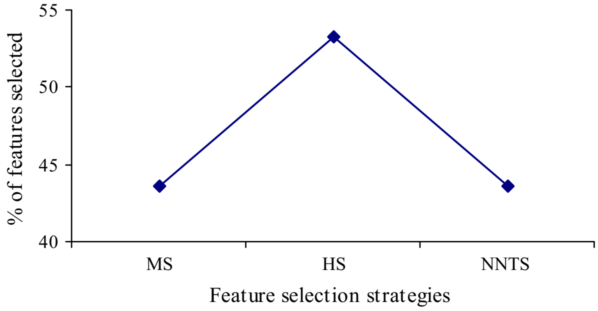

Fig. 7. Average % of features selected.

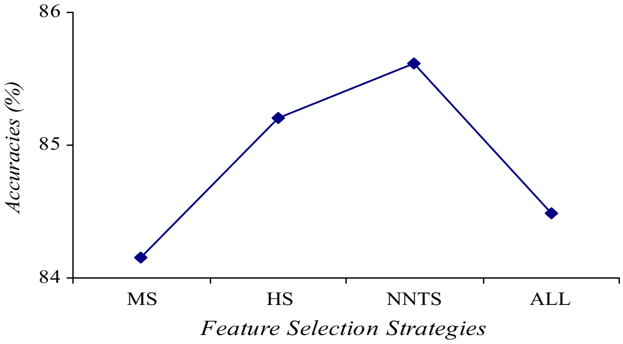

Fig. 8. Average accuracies.

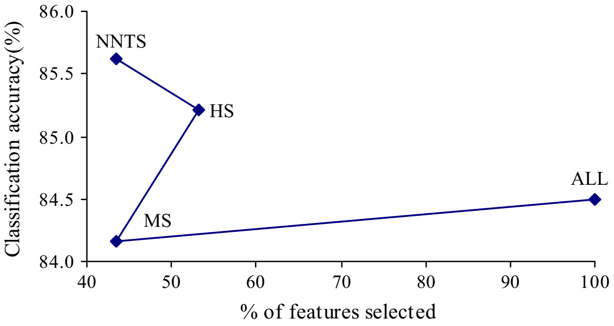

Fig. 9. Features selected (%) vs accuracy (%).

Additionally, Fig. 7 provides the relation between the average percentage of features selected and the feature selection strategies. Using the MS and NNTS strategies, an average of 43.5% of features is selected for classification. An average of 53.2% of features is selected by the HS strategy for classification. In all cases, the NNTS strategy performs well with a lower number of features and provides higher classification accuracy. Fig. 8 shows a plot between the average accuracies and the feature selection

Table 7
Comparison of classification accuracies (%) of RBF classifier using seven feature selection algorithms.

S.no.	Unselect	Fuzzy entropy-NNTS	IG	MIFS	mRMR	NMIFS	FCBF	CMIM
1.	95.85	**97.28**	96.42	96.56	**97.28**	**97.28**	96.85	**97.28**
2.	73.83	76.04	75.91	74.73	75.52	75.52	**76.17**	75.91
3.	84.44	**85.18**	84.81	83.70	84.44	84.81	82.96	83.33
4.	84.52	85.16	83.22	83.22	83.22	83.87	84.51	**85.8**
5.	83.82	**84.46**	82.43	76.35	83.78	84.12	80.74	83.78
Avg. H/E/L		85.62 2/1/2	84.59 0/0/5	82.91 0/0/5	84.85 0/1/4	85.12 0/1/4	84.25 1/0/4	85.22 1/1/3

strategies. The highest average accuracy of 85.62% is obtained by the NNTS strategy when compared with the MS and HS strategies. Fig. 9 shows the correlation between the average percentage of features selected for each feature selection strategy and the average classification accuracies. The MS strategy selected 43.5% of features to produce a classification accuracy of 84.16% and the HS strategy selected 53.2% of features to obtain a classification accuracy of 85.21%. The NNTS strategy achieves the highest classification accuracy of 85.62% with 43.5% of features selected.

Based on the analysis, the MS strategy selects those features whose fuzzy entropy value is greater than or equal to its mean value and the HS strategy selects the features that are placed in the top 50% when the features are sorted in descending order based on their fuzzy entropy values. In the above process, it is observed that some irrelevant features may be selected and that some relevant features may be omitted. However, the NNTS strategy facilitates retaining the most relevant features with an optimum reduction of features. Therefore, the fuzzy entropy NNTS is proven the best choice of feature selection strategy in the study of medical database classification.

Furthermore, the performance of the fuzzy entropy NNTS is compared with state-of-the-art feature selection algorithms, such as IG [7], MIFS [7], mRMR [8], NMIFS [9], FCBF [10] and CMIM [32], which are also shown in Table 7. The bold values in the entries are the highest among these seven feature selectors when used with the same classifier. The average accuracies of the same feature selectors are given in the row labeled 'Avg.'. The H/E/L row represents the number of higher, equal and lower accuracies obtained by each feature selection method for all the datasets under consideration. Table 7 also shows that the accuracy of the fuzzy entropy NNTS method is the same as the mRMR, NMIFS and CMIM feature selection algorithms, which are using the Breast Cancer dataset. Additionally, the results obtained by the fuzzy entropy NNTS method are proven empirically to be better than the other algorithms on the Heart-Statlog dataset and the Cleveland Heart Disease dataset. Thus, it is observed that the average accuracy of the proposed fuzzy entropy NNTS seems to be promising in the feature selection domain.

6. Conclusion

Feature selection aims to reduce the amount of unnecessary, irrelevant and redundant features. It helps retrieve the most relevant features in datasets and improves the classification accuracy with less computational effort. If the features are not chosen well, even the best classifier performs poorly. In this paper, we describe feature relevance measures based on fuzzy entropy values and devise three feature selection strategies: Mean Selection, Half Selection and Neural Network Threshold Selection with an RBF Network classifier. The intention is to select the correct set of features for classification when datasets contain noisy, redundant and vague information.

Five benchmark datasets from the UCI Machine Learning Repository are used for evaluation. The proposed feature selection strategies have produced accuracies that are acceptable or better when compared with the accuracy obtained for the entire feature set without any feature selection. Of all the proponents, the one that maximizes the accuracy is the fuzzy entropy with Neural Network Threshold Selection. In future, this can be applied to a wide range of problem domains with hybridization of different feature selection techniques to improve the performance of both the feature selection and the classification.

Conflict of interest statement

None declared.

References

[1] R.E. Abdel-Aal, GMDH based feature ranking and selection for improved classification of medical data, J. Biomed. Inform. 38 (6) (2005) 456–468.
[2] M.F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, Int. J. Expert Syst. Appl. 36 (2) (2009) 3240–3247.
[3] Chin-Yuan Fan, Pei-Chann Chang, Jyun-Jie Lin, J.C. Hsieh, A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification, Int. J. Appl. Soft Comput. 11 (1) (2011) 632–644.
[4] Huan Liu, Lei Yu, Toward integrating feature selection algorithms for classification and clustering, IEEE Trans. Knowl. Data Eng. 17 (4) (2005) 491–502.
[5] R. Kohavi, George H. John, Wrappers for feature selection subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.
[6] II-Seok Oh, Jin-Seon Lee, Byung-Ro Moon, Hybrid genetic algorithm for feature selection, IEEE Trans. Pattern Anal. Mach. Intell. 26 (11) (2004) 1424–1437.
[7] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Netw. 5 (4) (1994) 537–550.
[8] Hanchuan Peng, Fuhui Long, Chris Ding, Feature selection based on mutual information: criterion of max-dependency, max-relevance and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[9] Pablo A. Extevez, Michel Tesmer, Claudio A. Perez, Jacek M. Zurada, Normalized mutual information feature selection, IEEE Trans. Neural Netw. 20 (2) (2009) 189–201.
[10] Lei Yu, Huan Liu, Efficient feature selection via analysis of relevance and redundancy, J. Mach. Learn. Res. 5 (2004) 1205–1224.
[11] Seral Sahan, Kemal Polat, Halife Kodaz, Salih Gunes, A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis, Int. J. Comput. Biol. Med. 37 (3) (2007) 415–423.
[12] P. Jaganathan, K. Thangavel, A. Pethalakshmi, M. Karnan, Classification rule discovery with ant colony optimization and improved quick reduct algorithm, IAENG Int. J. Comput. Sci. 33 (1) (2007) 50–55.
[13] Rahul Karthik Sivagaminathan, Sreeram Ramakrishnan, A hybrid approach for feature subset selection using neural networks and ant colony optimization, Int. J. Expert Syst. Appl. 33 (1) (2007) 49–60.
[14] Shih-Wei Lin, Zne-Jung Lee, Shih-Chieh Chen, Tsung-Yuan Tseng, Parameter determination of support vector machine and feature selection using simulated annealing approach, Int. J. Appl. Soft Comput. 8 (4) (2008) 1505–1512.
[15] Shih-Wei Lin, Shih-Chieh Chen, Wen-Jie Wu, Chih-Hsien Chen, Parameter determination and feature selection for back-propagation network by particle swarm optimization, Int. J. Knowl. Inf. Syst. 21 (2) (2009) 249–266.
[16] Alper Unler, Alper Murat, A discrete particle swarm optimization method for feature selection in binary classification problems, Eur. J. Oper. Res. 206 (3) (2010) 528–539.
[17] Pei-Chann Chang, Jyun-Jie Lin, Chen-Hao Liu, An attribute weight assignment and particle swarm optimization algorithm for medical database classification, Int. J. Comput. Methods Progr. Biomed. 107 (3) (2012) 382–392.
[18] Maria Salamo, Maite Lopez-Sanchez, Rough set based approaches to feature selection for case-based reasoning classifiers, Int. J. Pattern Recognit. Lett. 32 (2) (2011) 280–292.

[19] Sultan Noman Qasem, Siti Mariyam Shamsuddin, Radial basis function network based on time variant multi-objective particle swarm optimization for medical diseases diagnosis, Int. J. Appl. Soft Comput. 11 (1) (2011))1427–1438.
[20] C E Shannon, A mathematical theory of communications, Bell Syst. Tech. J. 27 (379–423) (1948) 623–656.
[21] Bart Kosko, Fuzzy entropy and conditioning, Int. J. Inf. Sci. 40 (2) (1986) 165–174.
[22] H.M. Lee, C.M. Chen, J.M. Chen, Y.L. Jou, An efficient fuzzy classifier with feature selection based on fuzzy entropy, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 31 (3) (2001) 426–432.
[23] N.R. Pal, J.C. Bezdek, Measuring fuzzy uncertainty, IEEE Trans. Fuzzy Syst. 2 (2) (1994) 107–118.
[24] James C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer academic Publishers, Norwell, USA, 1981.
[25] D.K. Pratihar, Soft Computing, Narosa Publishing House, New Delhi, 2008.
[26] Rami N. Khushaba, Adel Al-Jumaily, Ahmed Al-Ani, Novel feature extraction method based on fuzzy entropy and wavelet packet transform for myoelectric control, in: ISCIT'07: International Symposium on Communications and Information Technologies, 17–19 October, 2007, Sydney, Australia, pp. 352–357.
[27] S. Rajasekaran, G.A. Vijayalakshmi Pai, Neural Networks, Fuzzy logic and Genetic Algorithms Synthesis and Applications, PHI Learning Pvt Ltd, New Delhi, India, 2010.
[28] Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur, Effective diagnosis of heart disease through neural networks ensembles, Int. J. Expert Syst. Appl. 36 (4) (2009) 7675–7680.
[29] A. Frank, A. Asuncion, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2010. 〈http://www.archive.ics.uci.edu/ml〉.
[30] Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 2011 〈http://www.cs.waikato.ac.nz/ml/weka〉. (Data Mining Software).
[31] Ron Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence-Vol. 2, 1995, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp. 1137–1143.
[32] F. Flueret, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.